Choosing the weights for the logarithmic pooling of probability distributions

Luiz Max F. de Carvalho[a,b,c], Daniel A. M. Villela[a], Flavio Coelho[c] & Leonardo S. Bastos[a]
a – Program for Scientific Computing (PROCC), Oswaldo Cruz Foundation.
b – Institute of Evolutionary Biology, University of Edinburgh.
c – School of Applied Mathematics, Getulio Vargas Foundation (FGV).

February 17, 2015

Abstract

Combining different prior distributions is an important issue in decision theory and Bayesian inference. Logarithmic pooling is a popular method to aggregate expert opinions by using a set of weights that reflect the reliability of each information source. The resulting pooled distribution however heavily depends set of weights given to each opinion/prior. In this paper we explore three objective approaches to assigning weights to opinions. Two methods are stated in terms of optimization problems and a third one uses a hierarchical prior that accounts for uncertainty on the weights. We explore an example in which a proportion is estimated using weights assigned to a finite set of opinions and show that, depending on the used method, results vary from discarding some of the expert opinions to the situation in which all opinions are assigned equal weights. Nevertheless, the three methods explored in this paper lead to very similar combined priors, with very similar integrated (marginal) likelihoods.

Key-words: logarithmic pooling; expert opinion; maximum entropy; Kullback-Liebler divergence; Dirichlet prior.

Background

Combining probability distributions is a topic of general interest, both in the statistical (Genest et al., 1986; Genest and Zidek, 1986) and decision theory literatures (Genest et al., 1984). On the theoretical front, studying opinion pooling operators may give important insights on consensus belief formation and group decision making (Genest and Zidek, 1986). Among the various opinion pooling operators proposed in the literature, logarithmic pooling has enjoyed much popularity, mainly due to its many desirable properties such as relative propensity consistency (RPC) and external Bayesianity (EB) (Genest et al., 1986). In a practical setting, logarithmic pooling finds use in a range of fields, from infectious disease modelling (Coelho and Codeço, 2009) and wildlife conservation (Poole and Raftery, 2000) to engineering (Lind and Nowak, 1988; Savchuk and Martz, 1994).

A common situation of interest is that of combining expert opinions, represented as proper probability distributions, about a quantity of interest $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n$. To combine these opinions using logarithmic pooling requires assigning weights to each of the experts. These weights represent the reliability of each opinion (Genest et al., 1984). This requirement naturally leads to the question of how to choose the weights in a meaningful fashion, according to some well-accepted optimality criterion. There are a few proposals in the literature that build methods using different approaches. One proposal is to maximise the entropy the pooled distribution (Myung et al., 1996), whereas another one is to minimise Kullback-Liebler (KL) divergence between the pooled distribution and the individual opinions (Abbas, 2009) or between the pooled (prior) distribution and the posterior distribution (Rufo et al., 2012a,b).

These approaches, while moving away from the problem of arbitrarily assigning the weights, arrive at single point solutions, similar to point estimates in Statistical theory. Albeit acknowledging that these approaches have merit, we argue that in many settings, where one has substantial prior information on the relative reliabilities of the information sources (experts), it would be desirable to incorporate this information into the pooling procedure while accommodating uncertainty about the weights (Poole and Raftery, 2000). Moreover, assigning a probability distribution to the weights permits us to obtain a posterior distribution using a Bayesian procedure, which in turn enables us to learn about these weights. Therefore, it makes possible to sequentially update knowledge about the reliability of each expert/source in the face of new data.

In this paper we explore previous approaches for deriving the weights for logarithmic pooling, namely by maximising the entropy of the resulting distribution and minimising the KL divergence between the pooled distribution and each individual distribution. Additionally, we propose a hierarchical prior approach in which we place a Dirichlet prior on the weights. We present an example on proportion estimation by combining Beta priors.

In what follows, we introduce the necessary theory and extend a previous result (Poole and Raftery, 2000) for combining more than two distributions.

Let $\mathbf{F}(\theta) = \{f_0(\theta), f_1(\theta), f_2(\theta), \ldots, f_K(\theta)\}$ be the set of prior distributions representing the opinions of $K+1$ experts and let $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_K\}$ be the vector of weights, such that $\alpha_i > 0 \; \forall i$ and $\sum_{i=0}^{K} \alpha_i = 1$. Then the log-pooled prior is

$$\pi(\theta) = t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} \tag{1}$$

where $t(\alpha) = \int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} d\theta$.

Logarithmic pooling will only yield proper probability distributions if it is possible to normalise the expression in (1). This condition is usually assumed implicitly, without proof. Poole and Raftery (2000) provide a proof for the case of two densities (see Theorem 1 therein), which we extend for the case of a finite number of densities.

Theorem 1. *Let A be the $(K+1)$-dimensional open simplex on $[0,1]$. For all $\boldsymbol{\alpha} \in A$ there exists a constant $t(\boldsymbol{\alpha})$ such that $\int_{\boldsymbol{\Theta}} \pi(\theta) d\theta = 1$.*

Here we provide a simple proof using Hölder's inequality.

Proof. We begin by noting that $\pi(\theta)$ can be re-written as:

$$\pi(\theta) \propto f_0(\theta) \prod_{j=1}^{K} \left(\frac{f_j(\theta)}{f_0(\theta)} \right)^{\alpha_j} \tag{2}$$

Let $X_j = \frac{f_j(\theta)}{f_0(\theta)}, j = 1, 2, \ldots, K$. Then integrating the expression in (2) is equivalent to finding

$$E_0 \left[\prod_{j=1}^{K} X_j^{\alpha_j} \right] \leq \prod_{j=1}^{K} E_0[X_j]^{\alpha_j} \tag{3}$$

where $E_0[\cdot]$ is the expectation w.r.t f_0 and (3) follows from Hölder's inequality for expectations (Yeh, 2011). Since we have, $\forall j$, $E_0[X_j]^{\alpha_j} = \left(\int_{\boldsymbol{\Theta}} f_0(\theta) \frac{f_j(\theta)}{f_0(\theta)} \right)^{\alpha_j} d\theta = 1^{\alpha_j} = 1$, Theorem 1 is proven. □

We now move on to study three approaches to assign weights, the first two approaches based on optimality criteria and a proposal based on pooling Dirichlet prior distributions.

Choosing the weights based on optimality criteria

Maximum entropy

In a context of near complete uncertainty about the relative reliabilities of the experts (information sources) it may be desirable to combine the prior distributions such that $\pi(\theta)$ is maximally uninformative. Such approach would ensure that, given the constraints imposed by $\mathbf{F}(\theta)$, the pooled distribution is the one which best represents the current state of knowledge (Jaynes, 1957; Savchuk and Martz, 1994). In order to choose $\boldsymbol{\alpha}$ so as to maximise prior diffuseness, one can maximise the entropy of the log-pooled prior:

$$H_\pi(\theta) = E_\pi\left[-\ln \pi(\theta)\right] = -\int_{\boldsymbol{\Theta}} \pi(\theta) \ln \pi(\theta) d\theta \tag{4}$$

In some cases it may be useful to express $H_\pi(\theta)$ as

$$H_\pi(\theta; \boldsymbol{\alpha}) = \sum_{i=0}^{K} \alpha_i E_\pi[-\ln f_i(\theta)] - \ln t(\boldsymbol{\alpha}) \tag{5}$$

Formally, we want to find $\hat{\boldsymbol{\alpha}}$ such that

$$\hat{\boldsymbol{\alpha}} := \arg\max H_\pi(\theta; \boldsymbol{\alpha}) \tag{6}$$

This approach, however, does not result in a convex optimisation problem, therefore one is not guaranteed to find a unique solution. See Proposition 1, below, for intuition as to why.

Minimising Kullback-Liebler divergence

One could also want to choose the pooling weights so as to minimise the total Kullback-Liebler divergence between each proposed distribution and the pooled distribution. Let $d_i = \mathrm{KL}(f_i||\pi)$ and let $L(\boldsymbol{\alpha})$ be a loss function such that

$$L(\boldsymbol{\alpha}) = \sum_{i=0}^{K} d_i \tag{7}$$

$$= -K \ln t(\boldsymbol{\alpha}) + \sum_{i=0}^{K}\sum_{j\neq i}^{K} \alpha_j \mathrm{KL}(f_i||f_j) \tag{8}$$

$$\hat{\boldsymbol{\alpha}} := \arg\min L(\boldsymbol{\alpha}) \tag{9}$$

Proposition 1. *The distribution obtained following (9) is unique, i.e., there is only one aggregated prior $\pi(\theta)$ that minimizes $L(\boldsymbol{\alpha})$.*

This property is proven in Rufo et al. (2012a). One can get some intuition into the proof of this claim by noting that minimising (8) is equivalent to maximising $\ln t(\boldsymbol{\alpha}) = \ln \int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} d\theta$. Rufo et al. (2012a) show that $t(\boldsymbol{\alpha})$ is concave, therefore the problem in (9) has a unique solution. By contrast, the problem in (6) requires to minimise $\ln t(\boldsymbol{\alpha})$ hence lacking a sufficient condition for the existence of a unique solution.

Specifying a prior distribution for $\boldsymbol{\alpha}$

In this section we propose a hierarchical prior for θ conditional on $\boldsymbol{\alpha}$ in order to incorporate uncertainty on the weights. A natural choice for a prior distribution for $\boldsymbol{\alpha}$ is the $(K+1)-$dimensional Dirichlet distribution. The conditional distribution $\pi(\theta|\boldsymbol{\alpha})$ is of the form in (1) and the prior density for $\boldsymbol{\alpha}$ is

$$\pi(\boldsymbol{\alpha}) = \frac{1}{\mathcal{B}(\boldsymbol{X})} \prod_{i=0}^{K} \alpha_i^{x_i - 1} \tag{10}$$

where $\boldsymbol{X} = \{x_0, x_1, \ldots, x_K\}$ is the vector of hyperparameters for the Dirichlet prior and $\mathcal{B}(X)$ is the multinomial beta function. The marginal prior for θ is then

$$\pi(\theta) = \int_A \pi(\theta|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \tag{11}$$

$$= \frac{1}{\mathcal{B}(\boldsymbol{X})} \int_A t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} \alpha_i^{x_i-1} d\boldsymbol{\alpha} \tag{12}$$

Application: binomial probabilities

We now turn our attention to combining expert opinions about probabilities and proportions. In this setting we are interested in the random variable $Y \sim \mathit{Bernoulli}(\theta)$. Again let us assume that we want to obtain a combined prior for a proportion θ. A common choice for $\mathbf{F}(\theta)$ is the Beta family of distributions:

$$f_i(\theta; a_i, b_i) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i b_i)} \theta^{a_i - 1}(1-\theta)^{b_i - 1}$$

The log-pooled prior is then

$$\pi(\theta) = \prod_{i=0}^{K} f_i(\theta; a_i, b_i)^{\alpha_i} \tag{13}$$

$$\propto \prod_{i=0}^{K} \left(\theta^{a_i-1}(1-\theta)^{b_i-1}\right)^{\alpha_i} \tag{14}$$

$$\propto \theta^{a^*-1}(1-\theta)^{b^*-1} \tag{15}$$

with $a^* = \sum_{i=0}^{K} \alpha_i a_i$ and $b^* = \sum_{i=0}^{K} \alpha_i b_i$. Again, (15) is the kernel of a Beta distribution with parameters a^* and b^*, thus

$$H_\pi(\theta) = \ln B(a^*, b^*) - (a^* - 1)\psi(a^*) - (b^* - 1)\psi(b^*) + (a^* + b^* - 2)\psi(a^* + b^*) \tag{16}$$

For the beta family of distributions, the KL divergence between $f_i(\theta)$ and $\pi(\theta)$ is

$$\begin{aligned} d_i = KL(f_i||\pi) = \ln\left(\frac{\mathcal{B}(a^*, b^*)}{\mathcal{B}(a_i, b_i)}\right) &+ (a_i - a^*)\psi(a_i) + (b_i - b^*)\psi(b_i) \\ &+ (a^* - a_i + b^* - b_i)\psi(a_i + b_i) \end{aligned} \tag{17}$$

The marginal prior for θ is

$$\pi(\theta) = \frac{1}{\mathcal{B}(X)} \int_A \frac{1}{\mathcal{B}(a^*, b^*)} \theta^{a^*-1}(1-\theta)^{b^*-1} \alpha_i^{x_i - 1} d\boldsymbol{\alpha} \tag{18}$$

which can also be efficiently approximated through Monte Carlo sampling. We provide a simple implementation using the Stan (Stan Development Team, 2014) probabilistic programming language at `https://github.com/maxbiostat/opinion_pooling`. R code for the methods, figures and tables presented in this paper can also be found at the above link.

Here we analyse an example proposed by Savchuk and Martz (1994) (also discussed in Rufo et al. (2012b)) in which four experts are required supply prior information about the survival probability of a certain unit for which there have been $y = 9$ successes out of $n = 10$ trials. The experts express their opinion as prior means for the survival probability, which Savchuk and Martz (1994) then use to construct prior distributions with maximum variance given the restriction on the means. From the vector of prior means $\mathbf{m} = \{m_0 = 0.95, m_1 = 0.80, m_2 = 0.90, m_3 = 0.70\}$, the authors obtain the parameters of the beta distributions for each expert, $\mathbf{a} = \{a_0 = 18.10, a_1 = 3.44, a_2 = 8.32, a_3 = 1.98\}$ and $\mathbf{b} = \{b_0 = 0.955, b_1 = 0.860, b_2 = 0.924, b_3 = 0.848\}$. The resulting prior densities are show in the top panel of Figure 1. To complete the analysis, we place a diffuse $Dirichlet(\boldsymbol{\alpha}|\boldsymbol{X})$ prior on $\boldsymbol{\alpha}$ with $X_i = 1/4\, \forall i$. Finally, we propose to compare the prior distributions representing the experts' opinions as well as the combined distributions obtained by the different approaches using the integrated (marginal) likelihood (Raftery et al. (2007), eq. 9), $l(y) = \int_0^1 f(\theta|x)\pi(\theta)d\theta$.

Results and avenues of future research

Table 1 lists the weights proposed by each method. Figure 1 shows the prior and posterior distributions in each of the methods and also the case in which we assign an equal weight $(1/K)$ to each opinion. It is interesting to note that maximum entropy suggests to discard all opinions but one, which effectively leads to the maximum entropy. Since $t(\boldsymbol{\alpha})$ is concave, we expect to find the maximum entropy given by the boundary conditions, which may lead to border points in the simplex. Minimising Kullback-Liebler divergence between each prior and the pooled prior leads to finding a unique solution but in this case also suggests to discard two of the opinions. By contrast, using a hierarchical Dirichlet prior for the weights gives rather different results from the first two methods in proposing almost equal weights to each of the opinions. One can get insight into these results by looking at the integrated likelihoods in Table 2 and the densities in Figure 1, we note that all three methods lead to similar pooled distributions. Note that the only distribution with a substantially different $l(y)$ is that of Expert 3, who gave a rather divergent mean for the survival probability $(m_3 = 0.70)$.

In conclusion, if the prior distributions (opinions) are not radically different, all three methods will probably lead to similar combined priors. Although this is the case for the simple univariate example presented, it remains to be seen if this is the case for high-dimensional θ under complex sampling distributions. As the results presented in this paper make clear, future research shall be focused on cases where there is substantial heterogeinity in the available opinions. Moreover, a sensitivity analysis for $\pi(\boldsymbol{\alpha})$ is desirable to understand how much we can lead about the experts reliabilities *a posteriori*.

Acknowledgements

The authors would like to thank Professor Adrian Raftery (University of Washington) for helpful suggestions. DAMV was supported in part by Capes under Capes/Cofecub project (N. 833/15). FCC is grateful to Fundação Getulio Vargas for funding during this project.

Table 1: Weights obtained using the three methods for the proportion estimation problem. [1] – Kullback-Liebler [2] – Posterior mean for $\boldsymbol{\alpha}$.

Method	α_0	α_1	α_2	α_3
Maximum entropy	0.00	1.00	0.00	0.00
Minimum KL[1] divergence	0.04	0.96	0.00	0.00
Hierarchical prior[2]	0.26	0.24	0.26	0.23

Table 2: Integrated likelihoods ($l(y)$) for the priors of each expert as well as the combined priors. [1] Calculated using the posterior mean of $\boldsymbol{\alpha}$

Expert priors		Pooled priors	
Expert 0	0.237	Equal weights	0.254
Expert 1	0.211	Maximum entropy	0.211
Expert 2	0.256	Minimum KL	0.223
Expert 3	0.163	Hierarchical[1]	0.255

References

Abbas, A. E. (2009). A Kullback-Leibler view of linear and log-linear pools. *Decision Analysis*, 6(1):25–37.

Coelho, F. C. and Codeço, C. T. (2009). Dynamic modeling of vaccinating behavior as a function of individual beliefs. *PLoS Comput. Biol.*, 5(7):e1000425.

Genest, C., McConway, K. J., and Schervish, M. J. (1986). Characterization of externally bayesian pooling operators. *The Annals of Statistics*, pages 487–501.

Genest, C., Weerahandi, S., and Zidek, J. V. (1984). Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17(1):61–70.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, pages 114–135.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. II. *Physical Review*, 108:171–190.

Lind, N. C. and Nowak, A. S. (1988). Pooling expert opinions on probability distributions. *Journal of engineering mechanics*, 114(2):328–341.

Myung, I. J., Ramamoorti, S., and Bailey Jr, A. D. (1996). Maximum entropy aggregation of expert predictions. *Management Science*, 42(10):1420–1436.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: the bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics*, pages 1–45. Oxford University Press.

Rufo, M., Martin, J., Pérez, C., et al. (2012a). Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach. *Bayesian Analysis*, 7(2):411–438.

Rufo, M. J., Pérez, C. J., Martín, J., et al. (2012b). A bayesian approach to aggregate experts' initial information. *Electronic Journal of Statistics*, 6:2362–2382.

Savchuk, V. P. and Martz, H. F. (1994). Bayes reliability estimation using multiple sources of prior information: binomial sampling. *Reliability, IEEE Transactions on*, 43(1):138–144.

Stan Development Team (2014). Stan: A c++ library for probability and sampling, version 2.6.0.

Yeh, C.-C. (2011). Hölder's inequality and related inequalities in probability. *International Journal of Artificial Life Research (IJALR)*, 2(1):54–61.

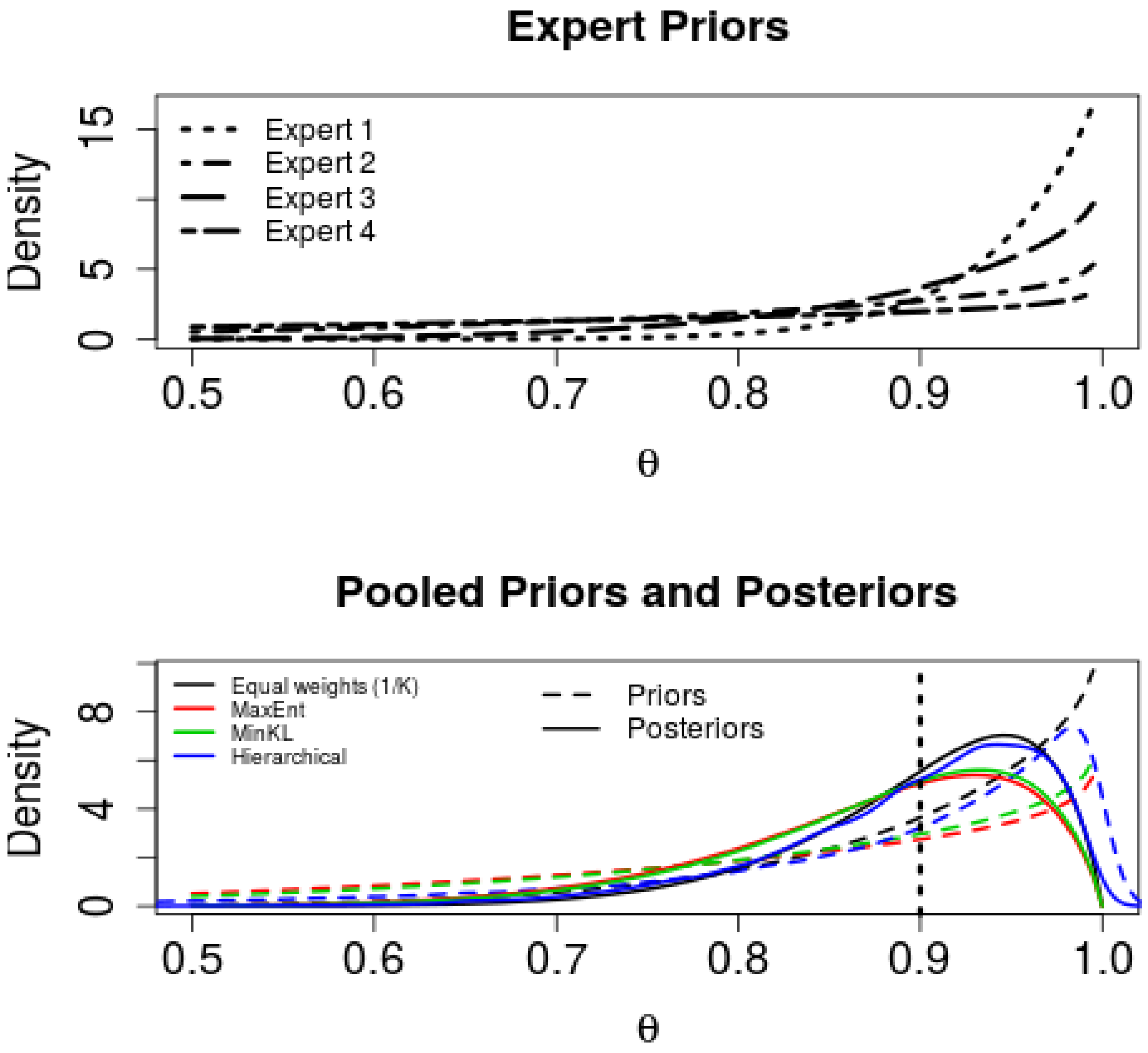

Figure 1: **Prior and posterior densities for** θ. Top panel shows the distributions elicited by each expert (data from Savchuk and Martz (1994)) and the bottom panel shows the pooled priors and posteriors obtained using each of the three methods discussed in this paper. The dashed vertical line marks the maximum likelihood estimate of θ, $\hat{\theta} = 9/10$.